# Data Analysis Report for Mingar's Product Management Team

Exploring Customer's Characteristics and Checking the Weakness of the Product

Report prepared for MINGAR by [Smart Analysis]

2022-04-11

# Contents

## Executive summary

Mingar, a company that produces high-end wearable fitness tracking devices, has recently launched new product lines to better compete with Bitfit, using a more approachable and affordable pricing strategy to attract customers.

### Research Question 1: Who are our customers? Are there differences between old and new customers?

This section we analyze the user profile of the users who bought the new product line and compare it with the user profile of the previous high-end product line, aiming to help Mingar locate the target customers of the new product line and allow Mingar to promote the new product in a more targeted way. The research results of the above focuses are shown below:

- The new product lines sell better than the traditional product lines, which is a positive signal that the affordable products are competitive.
- Both traditional and new consumers, the wearable fitness tracking devices are more popular among female than the male.
- The spread of new customers is wider between each feature, like age. More teenagers are likely to purchase this new product. However, the average age of the new customers is larger than the average age of the traditional customers. This also means that there will be more older consumers buying new products than traditional ones.

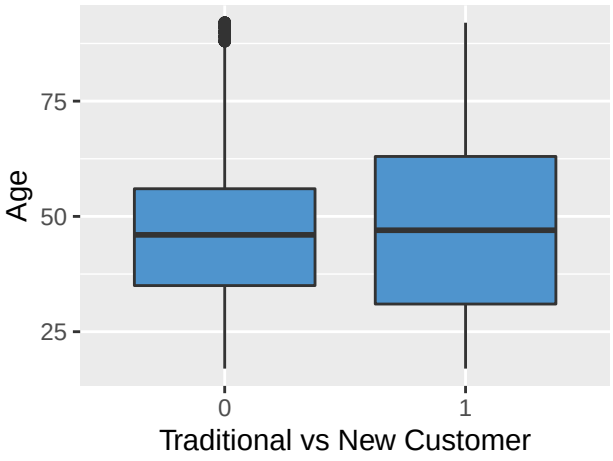


**Figure 1:** Comparsion of Age between Traditional and New Customer

- The average household median income for the new customers shows to be lower than the

traditional customers average household median income. The new product can target customers with lower income.

### Research Question 2: Are there performance differences related to users' skin color? Does it Varies Within Product lines?

Another research focus of this report is whether Mingar's product performance differs among users of different skin tones. Since Mingar is questioning whether their wearable devices were more likely to have abnormal or even missing data when monitoring the sleep of dark-skinned users, we conducted a correlation study between the sleep data and the skin color of the users. To reflect the company's value of equality, the data used by Smart Strategy does not contain any information that directly indicates the user's ethnicity or skin color. We inferred the user's likely skin color based on the color of the Emoji they used in the app, and analyzed it in conjunction with their sleep data records. The research results of the above focuses are shown below:

- Flag count reflects how frequent the device has abnormal or missing data. According to the study between the skin colors provided by emoji and the frequency of device reporting flags, the darker the skin color, the more frequently the device reports flags.

**Table 1:** Expected flag per 100 minutes of sleep for users with different skin colors, under the assumption that all users are 18 years old

| Skin Color | Reported Flags per 100 Minutes Sleep |
|---|---|
| Light | 0.31 |
| Medium-light | 0.68 |
| Medium | 1.01 |
| Medium-dark | 2.06 |
| Dark | 3.39 |

- For users with dark skin, the frequency of their devices reporting a flag is ten times higher than the frequency of the device reporting a flag for users with the lightest skin.
- The sleep monitoring quality of the devices is not related to the income of the user's neighborhood. We indicate that it is also not related to the user's income.
- For all product lines in Mingar, the performance of the devices on sleep scoring is poor and almost indifferent for users with dark skin.

# Technical report

## Introduction

The purpose of this report is to analyze the target market for the "active" and "advance" lines of Mingar's wearable fitness tracking devices based on consumer characteristics and to compare them with the characteristics of previous high-end product lines. Another focus of the study was the analysis of potential problems with product performance, especially regarding data anomalies recorded during sleep for dark-skinned users. The report begins by defining the research questions, presenting the data, which includes data description and data processing, and visualizing the data for clarity. The next section is a discussion of the results we found, that is, finding the variables that actually correlate with the two research questions. In the last section, the limitations and strengths of our work are shown, which is to explain some unexpected results and to improve them for the next time.

### Research question 1

- What type of customers would buy the new devices to find the features of our target customers?
- What are the differences between the new customers and the traditional customers?

### Research question 2

- Are the devices performing poorly for users with darker skin with respect to sleep scoring function?
- If there are some differences between the performance of the devices of the users with darker skin and lighter skin, does the performance vary within product lines?

### Data Description

The first part of the study focused on the characteristics about the customers, so the data we collated and used contained some basic information about each customer. For example, the age of the consumer is an important characteristic for the target market of the product, which can help us to make a clear positioning. And, the variables describing customers by gender provide differences in the purchasing power of new customers by gender. The median household income of the user's community is also included in the dataset as a way to determine and analyze the purchasing power of customers. Also, considering the privacy, the customer's neighborhood is recorded as Census subdivision unique identifier.

In the second second research, we focus on the sleep performance of darker skin users. Thus, the data we collected included the basic information and the sleeping data of customers as well. The most important variable is flags, which counts the occurrence of quality flags during a sleep period. Flags may appear because data is missing or there is a data quality condition such as a sensor error. Also, the variable recording the skin color of the user is just as important and is one of the focuses of our research. In addition, we use the customer ID in the dataset to identify the data provided by different users, which is a unique code. Additionally, duration records the time in minutes that the device records sleep. We use the data under this variable to calculate the number of flags that appear per minute for further comparison. What's more, age and community average income were also among the factors we analyzed, but they were rescaled in a 0-1 interval. Furthermore, we have also considered the difference between different products, so the lines of products the device belongs to is also included.

### Data Manipulation

Since we do not consider the case that the same customer will use more than one device, the original dataset, *customer* and *cust_dev* both record 19241 data, we combine the two data sets together according to the id of customers; and then we combine the new one and dataset of device, according to the id of devices, which form a new dataset *customer_output*.

We know from the postcode dataset that the same postcode will have multiple records (data of repeat postcodes), and the same postcode may also correspond to different Census subdivision unique identifiers so that the data of median income will be affected in the subsequent analysis. Therefore, the first Census subdivision unique identifier record that appears is used as the Census subdivision unique identifier corresponding to the postcode (the first one is selected when the duplicate value appears), in order to avoid multiple duplicates of postcode records when combining the data set. This allows us to merge the *postcode* with the *median_income* and count the population and income data corresponding to each postcode, forming a new dataset

*postcode_income*.

However, when merging the above two new datasets by the postcode, there are some duplicate and unwanted information, including postcode, pronouns, and id of device. In particular, the postcode may have privacy implications and is a duplicate of the Census subdivision unique identifier; the pronouns are also a duplicate of the sex and will not be considered; the id of device, similarly, is a unique id like Census subdivision unique identifier. What's more, after looking at the data about gender, there are 196 missing values, which we choose to remove; for intersex, it is kept because it accounts for more. For the variable of age, we derived the age of each customer by subtracting the year of this year and the year of the birthday recorded in the dataset, and overwriting it in the original age variable.

For the second research question, we need to count the skin color of users. Based on the comparison of skin color and modifier on the emoji website, we replaced the records for modifiers in the previously generated *customer_output* with the more understandable categories of "Dark", " Medium","Light", etc. However, some of the users did not modify the color of the emoji, so they were recorded under the default category. In addition, to further facilitate the subsequent analysis, we found that some variables are stored as"character" in the data set, which may not work in some models, so we changed them to "factor". Also, we merge the *cust_sleep* and *customer_output* according to the id of customer.

Then, because some predictor variables exist at different scales, we also rescale the age and median income variables to ensure that the model works properly, e.g., age equal to 0 corresponds to the youngest age of consumers in the dataset while age equal to 1 corresponds to the oldest age. Our goal is to investigate whether dark-skinned customers experience more problems when using the sleep monitoring feature of the Mingar device. The quantification taken in the data is the number of times the device flags during the sleep monitoring, but since each person does not sleep equal hours per day, a more reasonable approach is to compare how often the flags appear. Therefore, we added a variable to the data set to reflect the average number of flags per minute of sleep.

## Research Question #1: Who are our customers? Are there differences between old and new customers?

In order to broaden the market share, we first divided our products according to product lines in response to customer requests. Two of these products are the "Active" and "Advanced" lines, which serve as a more approachable price point for the average person. The target market of them is new customers, which is what we are aiming for in our research. We then use data visualization to look at and compare the characteristics of traditional consumers and new consumers.

### Exploratory Data Analysis

First of all, considering the characteristics of the products and the different consumption views of men and women, we think that gender may be one of the variables that need to be analyzed. So we drew a bar plot, as shown in the following figure, showing the gender distribution of traditional and new consumers. We removed 196 observations with missing values in gender in order to reduce the impact of missing values on the subsequent analysis.

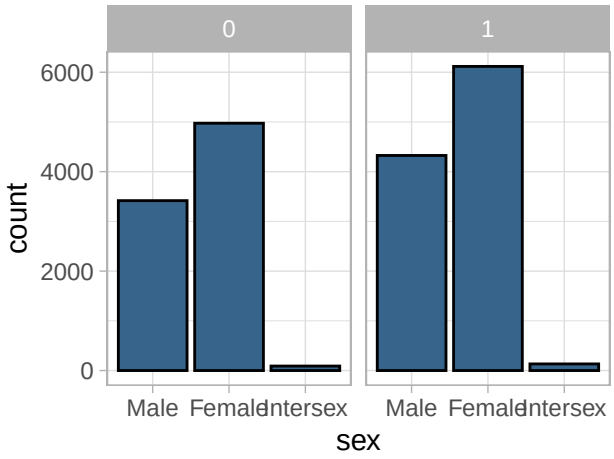


**Figure 2:** Comparison of distribution of gender between new and traditional customers

From the above plot, it is easy to see that the gender distribution of traditional and new consumers is relatively similar, with more female than male, and more female among the new consumers. So for the target market, gender is a characteristic worth considering, but not decisive.

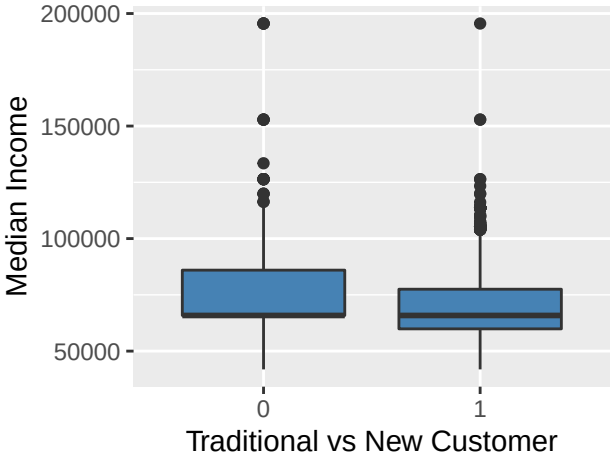


**Figure 3:** Comparsion of Median Income between Traditional and New Customer

Based on the box plot above depicting the distribution of median household income of area the customer lived between traditional and new clients, we find that the mean of median income of the neighborhoods for traditional and new clients is similar, both around $65,000 per year. The range is wider for traditional customers, but more downward for new customers, which also suggests that "active" and "premium" products are affordable. There are some outliers in the figure, but they are reasonable. Therefore we need to consider the median income of the community where the customer lives for the new consumer profile.

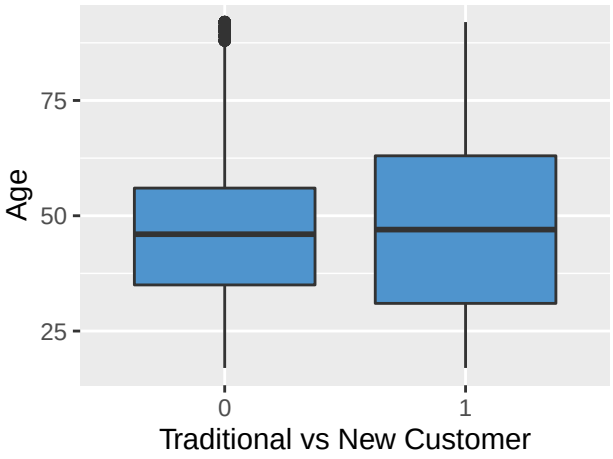


**Figure 4:** Comparsion of Age between Traditional and New Customer

The box plot above shows the age distribution of traditional and new customers. It is obvious that the range of new customers is broader, indicating that the new product can be accepted by a broader age group of consumers. However, their medians are similar, but slightly higher for new customers than for traditional customers. In this way, the age of the client also needs to be considered.

The table below shows the specific numbers for new and traditional clients. Also, it provides the comparative mean of age, household median income of the neighborhood they lived, and neighborhood population between new and traditional data to better demonstrate the actual difference.

From the table above, we know that the total number of new customers is greater than traditional customers. The median income of the neighborhoods the customers lived of the new customers are lower than the traditional ones. It is obvious and in line with our original target features for the new products, which is affordable for average people. However, new customers and traditional customers age by similar means, as we discussed previously. Furthermore, there is a gap in their mean of population of the neighborhoods they lived, but it does not seem to be significant.

From the above EDA, we can draw a preliminary conclusion that the sex, household median income, age, skin, and neighborhood population have some relationship between new and traditional customers. The modeling part is used to explore the actual effect from these factors.

### Modeling

To further study what are the outstanding factors for people becoming our new customer, we use generalized linear model with sex, age, skin and median income as fixed effect, and CSDuid (census subdivision unique identifier) as the random effect, and we fitted the model as:

$$Logit(p) = \beta_0 + \beta_1 * sex + \beta_2 * age + \beta_3 * skin + \beta_4 * median\ income + (1|CSDuid)$$

Here is a table of the results of the generalized linear model:

**Table 2:** Summary of the Generalized Linear Model

| Parameter | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| Intercept | 1.314e+00 | 8.386e-02 | 15.672 | < 2e-16 |
| sexFemale | -3.639e-02 | 3.028e-02 | -1.202 | 0.229 |
| sexIntersex | 1.406e-01 | 1.404e-01 | 1.001 | 0.317 |
| age | 5.053e-03 | 8.785e-04 | 5.752 | 8.82e-09 |
| skinDefault | -1.004e-02 | 5.029e-02 | -0.200 | 0.842 |
| skinMedium | 1.285e-02 | 5.695e-02 | 0.226 | 0.822 |
| skinMedium-light | 1.278e-02 | 5.622e-02 | 0.227 | 0.820 |
| skinLight | -3.824e-02 | 5.527e-02 | -0.692 | 0.489 |
| skinMedium-dark | -4.259e-02 | 5.628e-02 | -0.757 | 0.449 |
| hhld_median_inc | -1.701e-05 | 8.878e-07 | -19.156 | < 2e-16 |

We found that in the full model, the skin and sex of customers are not significant. However, some variables are significant but on vary different scales. Therefore, to better fit our reduced model, we modified the age and median household income within range [0, 1]. The closer the value to 1, means the customer has higher median income household and is older. In the reduced model, the fixed effects are age and household median income, and the random effect is still the ID that represents the living location of the customers. The model now becomes:

$$Logit(p) = \beta_0 + \beta_1 * age + \beta_2 * median\ income + (1|CSDuid)$$

Here is a table of the results of the reduced generalized linear model:

**Table 3:** Summary of the Reduced Generalized Linear Model

| Parameter | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| Intercept | 0.65951 | 0.09112 | 7.238 | 4.56e-13 |
| rescale_age | 0.37929 | 0.06586 | 5.759 | 8.44e-09 |
| rescale_income | -2.62460 | 0.36797 | -7.133 | 9.85e-13 |

Our final model is:

$$Logit(p) = 0.65951 + 0.37929 * rescale\ age - 2.62460 * rescale\ income + (1|CSDuid)$$

Both parameters have P value less than 0.001 in this model, indicating that we have strong evidence against the null hypothesis that the coefficient of the parameter is 0. The coefficient for the re-scaled age has a positive coefficient. When the age of the customer increases, he or she is more likely to buy Mingar's new products. The coefficient for the re-scaled median household income is negative. Customers with higher median household income is liee likely to purchase for Mingar's new product. This trend aligns with our previous analysis.

The results of comparing the features of new customers and traditional customers are shown in the plots in section of exploratory data analysis. In addition, we also perform separate linear models for age and median household income of their living neighborhoods to identify specific differences in age and median income between traditional and new consumers.

The first model for age is

$$age = \beta_0 + \beta_1 * Status$$

where the status indicates whether the target customer is a new customer.

Here is a table of the results of the linear regression model:

**Table 4:** Summary of the Reduced Generalized Linear Model

| Parameter | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| Intercept | 46.2708 | 0.1833 | 252.437 | $<$ 2e-16 |
| is_new | 1.4449 | 0.2460 | 5.873 | 4.36e-09 |

The results show that age has a significant influence on whether the customer is new. The average age for new customer is 1.4449 years older than the average age of customers that purchase the traditional product,

Considering the new customer has defined as the products in "Active" and "Advanced" product lines, we choose to use linear mixed model to eliminate the difference through adding a random effect of devices names.

In this way, the second model for median income is

$$median\ income = \beta_0 + \beta_1 * Status + (1|device_name)$$

where the status indicates whether the target customer is a new customer.

Here is a table of the results of the linear mixed model:

**Table 5:** Summary of the Reduced Generalized Linear Model

| Parameter | Estimate | Std. Error | t value |
|---|---|---|---|
| Intercept | 71816 | 1698 | 42.290 |
| is_new1 | -4091 | 2637 | -1.551 |

Similarly, in this model, we know that the household median income of neighborhoods where the customers lived has a significant influence on whether the customer is new. The median income of their neighborhood for the new customers is $4091 lower than it of customers that purchase the traditional product. This is also in line with our request to broaden our market share and change from high-end goods to products that are affordable to the general public.

In conclusion, the main differences between new and traditional consumers are income and age. In terms of median income of the areas that customer lived, the new consumers have lower median income than the traditional consumers. The age range of consumers who buy the new

product is much larger than that of the traditional product. After fitting the models, it also proves what we observed before, that new customers are older and have lower median income in their area of residence compared to traditional customers, indicating that the company's objective of launching new products to gain more market share was achieved. The new product not only expands the customer base, but also becomes more competitive in terms of price compared to the previous product.

## Question #2.1: Are the devices performing more poorly in sleep scoring for users with darker skin?

To investigate whether the device would perform poorly for darker-skinned users in terms of sleep scoring, we recorded the number of times the device showed abnormal data or missing data during sleep. To make it easier to compare and analyze, we changed it to numbers of the flags appearing per minute through the flags and duration. In order to avoid the device being labeled as "racist", which is not in line with our company values, we used the skin color of the selected emoji as a reference to further analyze the device.

### Exploratory Data Analysis

Table 1 below shows the average flag per minute for users in different skin colors. Dark skinned users have the highest average flag that, for every 100 minutes of sleep in dark skinned users, there are on average 3.3 flags reflecting that there are missing or unusual data. The table also shows that the lighter the skin color is, the less average flags the devices show. Users with light skin only receive 0.003 flag per minute on average, less than 1/10 of that for dark skinned users.

**Table 6:** Average Flag per Minute for Users in Different Skin Colors

| Skin | flag/minute |
|---|---|
| Dark | 0.0333971 |
| Default | 0.0065174 |
| Medium | 0.0099131 |
| Medium-light | 0.0066252 |
| Light | 0.0030656 |
| Medium-dark | 0.0201984 |

Figure 1 blow plots the device performance on sleep scoring across different skin color users. It is quite obvious that as the number of flag per minute increases, the portion of dark skinned users also increases.

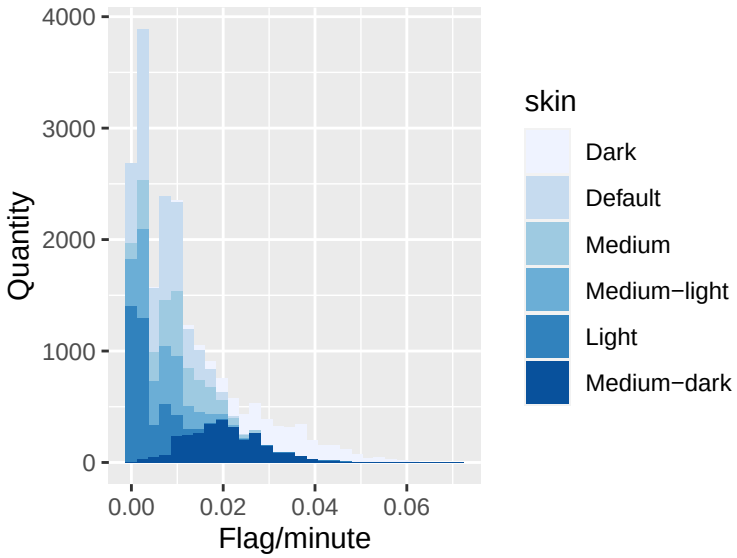


**Figure 5:** Device Performance Across Users With Different Skin Color Types

From the above EDA we draw a preliminary conclusion that dark skinned users are more likely to have data missing or abnormal data problems than users with lighter skin color. However, there may have other factors that may affect the frequency of the flag. To test our doubt, we want to explore the relationship between the frequency of flags and age.

Figure 2 shows that as the age of the users increases, there is a slightly downward trend of the frequency of flags. So there is a negative relationship between age and frequency of flags.

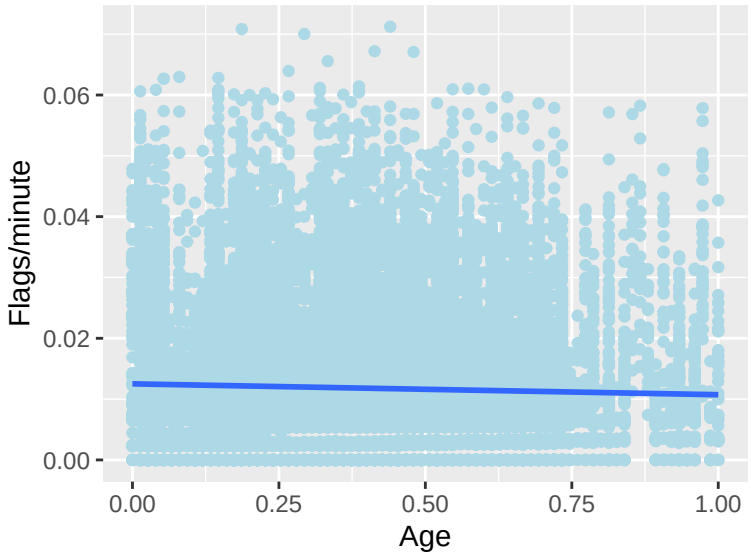


**Figure 6:** Relation Between Age and Average Flag/minute

### Modeling

Since we are modeling a count data, it is best to use Poisson regression to see if the log value of the observed value can be expressed by linear combination of some unknown parameters. We first fit a general linear mixed model with only skin color as the fixed effect, number of flags as response, and customer ID as random effect on intercept. An offset term of the duration of sleep in minutes is added in the model because we want to eliminate the effect caused by different sleep time for different observations and standardized the response.

The equation form of the first model is as the follow

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * skin + (1|customer\ id)$$

Considering other fixed effects on the number of flags, we fit the second model including skin color, age, and the household median income as the fixed effect, and the customer ID as the random effect. The offset term is still the time of sleep. Look into the coefficients of the general linear mixed model, there is one insignificant predictor: household median income. Including this predictor will not make the model fit better, so we fit another model without it.

The equation form of the second model is as the follow

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * skin + \beta_2 * age + \beta_3 * median\ income + (1|customer\ id)$$

The third model is to regress skin color and age on the number of flags, with a random variable of customer ID that affects the intercept under Poisson distribution. The offset term is the duration of sleep in minutes.

The equation form of the third model is as the follow

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * skin + \beta_2 * age + (1|customer\ id)$$

### Choose the Best Fit Model

Here is a table of the results of ANOVA:

**Table 7:** ANOVA Test Result

| Parameter | AIC | BIC | P-value |
|---|---|---|---|
| mod1 | 85563 | 85618 | |
| mod3 | 85557 | 85621 | 0.006874 |

We only perform an ANOVA test on the first and the third model. Considering the second model has an insignificant predictor compared to the third model and a larger AIC, the second model is not better than the third model. According to the ANOVA output, the complicated model has smaller AIC, larger BIC, and the p-value of the test is 0.0069. We have strong evidence against the null hypothesis that the simpler model is as good as the complicated model. So the best fit model is the third model.

Here is a table of the results of the Generalized linear mixed model:

**Table 8:** Summary of Final Model

| Parameter | Estimate | Std. Error | z value | P-value | 2.5% CI | 97.5% CI |
|---|---|---|---|---|---|---|
| Intercept | -3.383 | 0.009345 | -362.107 | < 2e-16 | -3.4024 | -3.3656 |
| skinDefault | -1.631 | 0.011888 | -137.235 | < 2e-16 | -1.6549 | -1.6082 |
| skinMedium | -1.212 | 0.012817 | -94.566 | < 2e-16 | -1.2373 | -1.1870 |
| skinMedium-light | -1.614 | 0.014043 | -114.905 | < 2e-16 | -1.6413 | -1.5861 |
| skinLight | -2.390 | 0.017450 | -136.966 | < 2e-16 | -2.4244 | -2.3559 |
| skinMedium-dark | -0.499 | 0.010935 | -45.675 | < 2e-16 | -0.5210 | -0.4780 |

| Parameter | Estimate | Std. Error | z value | P-value | 2.5% CI | 97.5% CI |
|---|---|---|---|---|---|---|
| age | -0.046881 | 0.0172 | -2.711 | 0.00671 | -0.0808 | -0.0129 |

Our final model is:

$$log(\frac{p}{1-p}) = -3.383919 - 1.631511 * skinDefault - 1.212081 * skinMedium-$$

$1.613651 * skinMedium-light-2.390047 * skinLight-0.499462 * skinMedium-dark-0.046881 * age$

The coefficients of all predictors are negative, which means comparing to the intercept, the existence of other factors indicates less frequency of flags. The intercept of -3.383921 means, for user who is 18 year-old, he/she will have exp(-3.383921) = 0.033914 flag per minute on average if he/she has dark skin. The predictor of users with light skin has the smallest coefficient of -2.390049, which means the frequency of flag for a 18 year-old user with light skin is exp(-2.390049) = 0.091625, about 9% of the frequency of users with dark skin at the same age. The coefficient of age is -0.046879. Since we rescaled the age to range it between 0 and 1, we can conclude there is a negative trend between age and the frequency of flag. The eldest users would have 1-exp(-0.046879) *100% = 4.58% less flags than the youngest users with the same skin color on average.

Based on the best fit model we build, we find out that devices do perform worse for users with dark skin. We also find out that, the darker the skin color, the more frequently the device report flags.

## Question #2.2: Does the Performance of the Device Varies Within Product lines?

In the previous research question, we found that the devices do perform more poorly on sleep scoring for users in dark skin. We want to further explore if the performance of devices for dark skinned users varies for different product lines, or if devices from all product lines perform just as poorly.

### Data Manipulation

To study the performance of the devices for users with dark skin, we need to filter out the data of only dark-skinned users from the data set that obtains sleep data.

### Modeling

Fit a general linear mixed model of product line on number of flags reported, with random effect of customer ID and offset term of duration of sleep. The summary of the coefficient of the fixed effect is showed below:

Here is a table of the results of the Generalized linear mixed model:

**Table 9:** The Results of the Generalized linear mixed model:

| Parameter | Estimate | Std. Error | z value | P-value |
|---|---|---|---|---|
| Intercept | -3.382480 | 0.009986 | -338.731 | <2e-16 |
| lineAdvance | -0.026341 | 0.013487 | -1.953 | 0.0508 |
| lineActive | -0.036850 | 0.022262 | -1.655 | 0.0979 |

There is weak evidence against the null hypothesis that the frequency of flags on devices from "advance" product line is the same as the frequency of flags on devices from "run" product line, and the predictor of the device from the "active" product line is not statistical significant. Based on this model, we conclude that there is little or no difference between the performance of devices from different product lines on monitoring the sleeping quality for dark-skinned users. All product lines need update to improve the monitoring quality for dark-skinned users.

## Discussion

The first research question was to study the customers profile of Mingar's "Active" and "Advanced" products and the differences between them and traditional customers in order to gain more market share in subsequent sales. The results of the exploratory data analysis showed that both old and new customers are more female; the average age difference between old and new customers is small, but the age range of the new customers was significantly wider; and the new customers were slightly lower in median income than the traditional customers.

The initial model is a generalized linear mixed model which includes the age, skin and household median income. Since area code is set as a random variable. The result shows that only the age and household median income of the customers have p-values that are less than 0.05. We then refitted the model after omitting the insignificant variables, with only age and household median income as fixed effect left. Both predictors are significant when age shows a positive correlation with customers buying new product but median household shows a negative correlation

with customers buying new product. This modeling result means that Mingar's new product is attracting people at older age and its strategy of aiming at providing affordable wearable electronic devices to more customers seems to be successful. Then, we use linear model, and linear mixed model to check the relationship between customers who buy new product and their age, and between customer who buy new product and their median household income. Both predictors in two models have significant P-values.On average, the mean age of Mingar's customers buying new product is 1.4449 years older than the mean age of the customers buying traditional products. The median household income for customers who buy Mingar's new products is $4091 lower than that of customers who buy the traditional products.

In conclusion, age and household median income are two variables that show the differences between new customers and traditional customers. The people with a higher age than the traditional customer are more attractive by the new product. Also, the new product becomes more attractive to the people who have a lower household median income obviously because it is more affordable.

The second research question focuses on the performance of Mingar's devices on sleep scores, as Mingar reported a trend in complaints on sleep scores that its devices performed poorly for users with darker skin. To avoid directly collecting information related to race, we collected the chat features that each user uses in his or her app and tagged the skin color of the emoji he/she uses as an indicator of his/her skin color. The result of the exploratory data analysis shows that users who are indicated to have dark skin have the most flag per minute of sleep, which is 0.033 flag per minute, and users who are indicated to have light skin have the least average flag per minute of sleep, which is 0.0031 flag per minute.

We then fit three general linear mixed models. By comparing the AICs and BIC between each model and then performing an ANOVA test, we choose the best fit model. All coefficients of the predictors on skin color of the model are negative and statistically significant, which means that compared to users with dark skin, all users with other skin colors have less flag per unit, if both have the same age. The negative predictor of age shows that, older people are likely to have less flag per unit, if they have the same skin color. The result of the model aligns with the result of the EDA. Another finding from the model is that, the lighter the skin color is, the more negative its coefficient is. As more negative coefficient leads to greater decrease in number of flags per unit, the model proves that Mingar's devices do perform more poorly on censoring sleep scores when the skin color of the user is darker. The darker the skin color is, the more poorly the device performs.

After proving there does exist performance differences between users with different skin colors, we want to test if Mingar's device from all its product lines performs as poorly when users have dark skin. By filtering data from users with dark skin and fitting a general linear model of the product

line on average flags per minute, we observe little differences across product lines. Although the coefficient of the "advance" product line shows weak evidence against the null hypothesis that the performance of devices from the "run" product line have the same performance as devices from the "advanced" product line, the coefficient of -0.02633 only reflects about 2.6% (1-exp(-0.02633)) difference of the performances between those two product lines.

In conclusion of research question 2, the devices performed more poorly on sleep scoring for users with darker skin color than users with lighter skin and devices from all product lines do not show much differences. We suggest that Mingar make improvements on the performance of sleep scoring of all its devices from all product lines to improve user experiences.

### Strengths and limitations

**Strengths** Before considering the variables that fit into the model, we use data visualization operations to analyze the data to determine if there is a correlation between the data and the response variable we want to analyze. In this way, we can avoid the complexity of model selection by adding variables that are not significant in the model. In addition, data visualization gives the reader a more intuitive way to understand the distribution and characteristics of the data than direct narration. The extension of the linear model we used can be used to do inference in addition to making predictions based on the confidence intervals, significance tests, and more.

**Limitations** There are some limitations due to the methods that are used. Because of the large data, the generalized linear mixed model could not handle all variables at the same time. Thus, some variables that have small effects may not be included in the model. When dealing with the data stored in the postal code and Census subdivision unique identifier, the data in these two variables are not one-to-one, and we choose to save only the first Census subdivision unique identifier that appears in the duplicate information corresponding to the postal code, but this still causes some data loss. Besides, in compiling the data provided on the website for emoji skin color selection, we assume that consumers with this skin color will use emoji with the same skin color, but most of them do not make a color selection and use the default color, and we cannot be sure that there will be people with this skin color who do not use this emoji. This may have an impact on our findings in the second study.

# Consultant information

## Consultant profiles

**Yichun Zhang** is a senior consultant with Smart Analysis. She specializes in data visualization and statistical modeling. Yichun earned her Bachelor of Science, Specialist in Statistics, from the University of Toronto in 2023.

**Hanqi Cui** is a senior consultant with Smart Analysis. She specializes in product management and data analysis. Hanqi earned her Bachelor of Commerce, Specialist in Finance and major in Statistics, from the University of Toronto in 2022.

**Weiyang Li**. Weiyang is a junior consultant with Smart Analysis. She specializes in reproducible analysis and statistical communication. Weiyang earned her Bachelor of Science, Majoring in Statistics and Cinema Studies from the University of Toronto in 2023.

**Yaqi Mu** is a junior consultant with Smart Analysis. She specializes in data analysis. Yaqi earned her Bachelor of Science, Specialist in Finance and Statistics, from the University of Toronto in 2022.

## Code of ethical conduct

1. We have complied with the norm of protecting human rights. When dealing with datasets that contain consumer postal codes, we chose to remove this variable and use the Census subdivision unique identifier instead, given the privacy concerns.
2. Our research is carried out and documented in a discreet manner, based solely on the client's requirements. No misleading data summaries are used, and the data and analysis of the feedback are true and valid. Furthermore, in order to avoid conflicts of personal interest or third party interests, we keep the content of the data used absolutely confidential and do not disclose any raw data.
3. The data we use is obtained through reasonable and permissible means. For example, through an open API on the website or by downloading from the library using a University of Toronto student ID.
4. We uphold professionalism and make analyses that are in line with industry development norms. All research analyses can be reproduced for review.

## References

[1] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[4] Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest

[5] Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

[6] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

[7] "Fitness Tracker Info Hub: Industry Standard Data for Wearable Fitness Trackers in the Candian Market." Fitness tracker info hub. Accessed April 4, 2022. https://fitnesstrackerinfohub.netlify.app/.

[8] "Postal Code Conversion File: 2016 Census Geography." Postal code conversion file: 2016 census geography | Map and Data Library, September 2021. Accessed April 4, 2022. https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file/2016.

[9] "Population Density." Census Mapper, February 9, 2022. Accessed April 4, 2022. https://censusmapper.ca/.

[10] von Bergmann, J., Dmitry Shkolnik, and Aaron Jacobs (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. v0.4.2.

[11] Hadley Wickham and Evan Miller (2021). haven: Import and Export "SPSS", "Stata" and "SAS" Files. R package version 2.4.3. https://CRAN.R-project.org/package=haven

[12] Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

[13] "Full Emoji Modifier Sequences, V14.0 - Unicode." Unicode: Emoji Charts. Accessed April 6, 2022. https://www.unicode.org/emoji/charts/full-emoji-modifiers.html.

[14] "Code of Ethical Statistical Practice." Statistical Society of Canada. Accessed April 7, 2022. https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf.

[15] Hadley Wickham, Jim Hester and Jennifer Bryan (2022). readr: Read Rectangular Text Data. R package version 2.1.2. https://CRAN.R-project.org/package=readr

[16] Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36

[17] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

[18] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

[19] Alboukadel Kassambara (2020). ggpubr: "ggplot2" Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

# Appendix

### Web scraping industry data on fitness tracker devices

First we need to determine whether it is reasonable to web scrape data with "bow" function, through the fitness tracker info hub to provide the data URL. After we get a response allowing scraping, we need to provide the information of the user agent and the output of the command shows the data is scrapable. After getting the scrapable response with some limitation of web scraping and determining the cautions, the target data we need within the web page will be extracted and imported into a table saved locally.

```
# Store the url of device data for web scraping
url <- "https://fitnesstrackerinfohub.netlify.app/"

# Add informative user_agent details
target <- bow(url,
              user_agent = "kimberlyzyc.zhang@mail.utoronto.ca for STA303/1002
              ↪  project",
              force = TRUE)

# Any details provided in the robots text on crawl delays and
# which agents are allowed to scrape
target

html <- scrape(target)
```

### Web scraping full-emoji-modifiers data on unicode

After storing the url of the website that contains the emoji we want to scrap, we first check if the data on the website is scrapable using the "bow" function. Then, we provide the information of the user agent and get the output of scrapping allowance and some restrictions on scrapping. After determining the cautions, we scrap the target data from the website and extract the emoji data we need into a table saved locally.

```
# Store the url of emoji data for web scraping
url <- "https://unicode.org/emoji/charts/full-emoji-modifiers.html"

# Add informative user_agent details
target <- bow(url,
              user_agent = "kimberlyzyc.zhang@mail.utoronto.ca for STA303/1002
              ↪  project",
```

```
              force = TRUE)

# Any details provided in the robots text on crawl delays and
# which agents are allowed to scrape
target

html <- scrape(target)
```

### Accessing Census data on median household income

On the census mapper website, we firstly register an account with our name and email. Then, we can get the API key in our profile. We need to install and activate the "cancensus" package in R, which is the package included the dataset we need. Then, we use the "options" function to get the regional data for the 2016 census with the API key. Finally we do some processing of the data to ensure that only the data we need is retained.

```
# Extract data through api key
options(cancensus.api_key = "",
        cancensus.cache_path = "cache") # this sets a folder for your cache

# get all regions as at the 2016 Census (2020 not up yet)
regions <- list_census_regions(dataset = "CA16")

regions_filtered <-  regions %>%
  filter(level == "CSD") %>% # Figure out what CSD means in Census data
  as_census_region_list()

# This can take a while
# We want to get household median income
census_data_csd <- get_census(dataset='CA16', regions = regions_filtered,
                              vectors=c("v_CA16_2397"),
                              level='CSD', geo_format = "sf")
```

### Accessing postcode conversion files

Using our University of Toronto student account, accept a license agreement in the online library and then download the data of postcode conversion in SPSS format.

```
# Load the postal code data
dataset = read_sav("data-raw/pccfNat_fccpNat_082021sav.sav")
```